Zeeman mapping of the magnetic field over extended regions (*30*).

References and Notes

1. A. G. W. Cameron, *Publ. Astron. Soc. Pac.* **69**, 201 (1957).
2. C. F. McKee, J. P. Ostriker, *Astrophys. J.* **218**, 148 (1977).
3. C. F. McKee, L. L. Cowie, J. P. Ostriker, *Astrophys. J.* **219**, L23 (1978).
4. T. W. Hartquist, *Astrophys. Space Sci.* **216**, 185 (1994).
5. K. W. Weiler, R. A. Sramek, *Annu. Rev. Astron. Astrophys.* **26**, 295 (1988).
6. J. Rho, R. Petre, *Astrophys. J.* **503**, L167 (1998).
7. R. A. Chevalier, *Astrophys. J.* **511**, 798 (1999).
8. D. P. Cox *et al.*, *Astrophys. J.* **524**, 179 (1999).
9. R. L. Shelton *et al.*, *Astrophys. J.* **524**, 192 (1999).
10. K. J. Borkowski, J. Rho, S. P. Reynolds, K. K. Dyer, *Astrophys. J.* **550**, 334 (2001).
11. D. A. Frail, W. M. Goss, V. I. Slysh, *Astrophys. J.* **424**, L111 (1994).
12. D. A. Frail *et al.*, *Astron. J.* **111**, 1651 (1996).
13. A. J. Green, D. A. Frail, W. M. Goss, R. Otrupcek, *Astron. J.* **114**, 2058 (1997).
14. B. Koralesky, D. A. Frail, W. M. Goss, M. J. Claussen, A. J. Green, *Astron. J.* **116**, 1323 (1998).
15. F. Yusef-Zadeh, D. A. Roberts, W. M. Goss, D. A. Frail, A. J. Green, *Astrophys. J.* **512**, 230 (1999).
16. M. Elitzur, *Annu. Rev. Astron. Astrophys.* **30**, 75 (1992).
17. M. Gray, *Philos. Trans. R. Soc. London Ser. A* **357**, 3277 (1999).
18. S. Weinreb, A. H. Barrett, M. L. Meeks, J. C. Henry, *Nature* **208**, 29 (1963).
19. H. Weaver, D. Williams, N. Dieter, W. Lum, *Nature* **208**, 29 (1965).
20. M. J. Reid, J. M. Moran, *Annu. Rev. Astron. Astrophys.* **19**, 231 (1981).
21. M. Miyoshi *et al.*, *Nature* **373**, 127 (1995).
22. M. Elitzur, *Astrophys. J.* **457**, 415 (1996).
23. W. D. Watson, H. W. Wyld, *Astrophys. J.* **558**, L55 (2001).
24. M. Elitzur, *Astrophys. J.* **504**, 390 (1998).
25. W. M. Goss, B. J. Robinson, *Astrophys. Lett.* **2**, 81 (1968).
26. M. Elitzur, *Astrophys. J.* **203**, 124 (1976).
27. L. K. DeNoyer, *Astrophys. J.* **228**, L41 (1979).
28. D. A. Frail, G. F. Mitchell, *Astrophys. J.* **508**, 690 (1998).
29. Y. Arikawa, K. Tatematsu, Y. Sekimoto, T. Takahashi, *Publ. Astron. Soc. Jpn.* **51**, L7 (1999).
30. F. Yusef-Zadeh, W. M. Goss, D. A. Roberts, B. Robinson, D. A. Frail, *Astrophys. J.* **527**, 172 (1999).
31. F. Yusef-Zadeh, F. Melia, M. Wardle, *Science* **287**, 85 (2000).
32. F. Yusef-Zadeh, S. R. Stolovy, M. Burton, M. Wardle, M. C. B. Ashley, *Astrophys. J.* **560**, 749 (2001).
33. M. J. Claussen, D. A. Frail, W. M. Goss, R. A. Gaume, *Astrophys. J.* **489**, 143 (1997).
34. C. L. Brogan, D. A. Frail, W. M. Goss, T. H. Troland, *Astrophys. J.* **537**, 875 (2000).
35. T. L. Bourke, P. C. Myers, G. Robinson, A. R. Hyland, *Astrophys. J.* **554**, 916 (2001).
36. P. Lockett, E. Gauthier, M. Elitzur, *Astrophys. J.* **511**, 235 (1999).
37. B. T. Draine, W. G. Roberge, A. Dalgarno, *Astrophys. J.* **264**, 485 (1983).
38. M. J. Kaufman, D. A. Neufeld, *Astrophys. J.* **456**, 611 (1996).
39. T. W. Hartquist, A. Sternberg, *Mon. Not. R. Astron. Soc.* **248**, 48 (1991).
40. T. W. Hartquist, K. M. Menten, S. Lepp, A. Dalgarno, *Mon. Not. R. Astron. Soc.* **272**, 184 (1995).
41. M. Wardle, F. Yusef-Zadeh, T. R. Geballe, in *The Central Parsecs of the Galaxy*, H. Falcke, A. Cotera, W. Duschl, F. Melia, M. Rieke, Eds. (Astronomical Society of the Pacific, San Francisco, 1999), p. 432.
42. M. Wardle, *Astrophys. J.* **525**, L101 (1999).
43. S. S. Prasad, S. P. Tarafdar, *Astrophys. J.* **267**, 603 (1983).
44. G. Pineau des Forêts, E. Roueff, D. R. Flower, *Mon. Not. R. Astron. Soc.* **223**, 743 (1986).
45. B. T. Draine, W. G. Roberge, *Astrophys. J.* **259**, L91 (1982).
46. D. F. Chernoff, C. F. Mckee, D. J. Hollenbach, *Astrophys. J.* **259**, L97 (1982).
47. M. G. Burton, D. J. Hollenbach, M. R. Haas, E. F. Erickson, *Astrophys. J.* **355**, 158 (1990).
48. M. J. Richter, J. R. Graham, G. S. Wright, D. M. Kelly, J. H. Lacy, *Astrophys. J.* **449**, L83 (1995).
49. M. I. Pastchenko, V. I. Slysh, *Astron. Astrophys.* **35**, 153 (1974).
50. A. J. Green, M. Wardle, J. S. Lazendic, paper presented at the 24th meeting of the IAU, Joint Discussion 1, Atomic and Molecular Data for Astrophysics: New Developments, Case Studies and Future Needs, Manchester, England, August 2000.
51. F. Yusef-Zadeh, K. I. Uchida, D. Roberts, *Science* **270**, 1801 (1995).
52. A. D. Gray, J. Nicholls, R. D. Ekers, L. E. Cram, *Astrophys. J.* **448**, 164 (1995).
53. J. S. Lazendic *et al.*, *Mon. Not. R. Astron. Soc.* **331**, 537 (2002).
54. J. S. Lazendic *et al.*, in preparation.
55. P. A. Shaver *et al.*, *Astron. Astrophys.* **147**, L23 (1985).
56. R. T. Stewart, R. F. Haynes, A. D. Gray, W. Reich, *Astrophys. J.* **432**, L39 (1994).
57. J. S. Lazendic *et al.*, in preparation.
58. D. J. Mullan, *Mon. Not. R. Astron. Soc.* **153**, 145 (1971).
59. P. W. J. L. Brand *et al.*, *Astrophys. J.* **334**, L103 (1988).
60. B. T. Draine, C. F. McKee, *Annu. Rev. Astron. Astrophys.* **31**, 373 (1993).
61. E. G. Zweibel, *Phys. Plasmas* **6**, 1725 (1999).
62. G. M. Dubner, P. F. Velásquez, W. M. Goss, M. A. Holdaway, *Astron. J.* **120**, 1933 (2000).

63. We thank G. Dubner, J. Hewitt, and J. Rho for assistance with Fig. 2 and J. Lazendic for assistance with Figs. 3 and 4. F. Y.-Z. acknowledges support by NASA grant NAG-5-9188.

REVIEW: AIDS

Diversity Considerations in HIV-1 Vaccine Selection

Brian Gaschen,[1] Jesse Taylor,[1] Karina Yusim,[1] Brian Foley,[1] Feng Gao,[2] Dorothy Lang,[1] Vladimir Novitsky,[3] Barton Haynes,[2] Beatrice H. Hahn,[4] Tanmoy Bhattacharya,[1] Bette Korber[1,5]*

Globally, human immunodeficiency virus–type 1 (HIV-1) is extraordinarily variable, and this diversity poses a major obstacle to AIDS vaccine development. Currently, candidate vaccines are derived from isolates, with the hope that they will be sufficiently cross-reactive to protect against circulating viruses. This may be overly optimistic, however, given that HIV-1 envelope proteins can differ in more than 30% of their amino acids. To contend with the diversity, country-specific vaccines are being considered, but evolutionary relationships may be more useful than regional considerations. Consensus or ancestor sequences could be used in vaccine design to minimize the genetic differences between vaccine strains and contemporary isolates, effectively reducing the extent of diversity by half.

Since HIV-1 M group began its expansion in humans roughly 70 years ago (*1*, *2*) it has diversified rapidly (*3*), now comprising a number of different subtypes and circulating recombinant forms (CRFs). The HIV-1 M group is the set of diverse viruses that dominates the global AIDS epidemic. Subtypes are genetically defined lineages that can be resolved through phylogenetic analysis of the HIV-1 M group as well-defined clades, or branches, in a tree. Recombination occurs frequently, and a CRF carries sections of two or more subtypes in a mosaic genome; a recombinant lineage is designated a CRF when related forms are found in multiple epidemiologically unlinked individuals. Currently, strains belonging to the same subtype can differ by up to 20% in their envelope proteins, and between-subtype distances can soar to 35%. Moreover this diversity is continually growing. The need for frequent changes in the annual influenza vaccine puts into perspective the implications of such diversity—less than 2% amino acid change can cause a failure in the cross-reactivity of the polyclonal response to the influenza vaccine and necessitates changing the vaccine strain (*4*).

[1]Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [2]Duke University AIDS Center, Durham, NC 27710, USA. [3]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115, USA. [4]University of Alabama at Birmingham, Birmingham, AL 35294, USA. [5]Santa Fe Institute, Santa Fe, NM 87501, USA.

*To whom correspondence should be addressed. E-mail: btk@t10.lanl.gov

Although the scale of the HIV-1 pandemic makes action imperative, there is still much to learn about the extent and immunological implications of HIV-1 sequence diversity. We do, however, have in hand the fruits of an extensive global HIV sequencing effort [currently there are 72,221 HIV sequences in the database (*3*)] that can provide a framework for reasoned vaccine strain selection. Optimizing selection is of the utmost urgency, as a number of human vaccine trials are being planned and initiated (*5*, *6*), and it is difficult to change strains during the long course of vaccine development, from initial concept to human trial. Subtype C is the most prevalent HIV-1 subtype globally, and it predominates in several geographic regions where vaccines might be evaluated. In these regions, the epidemiologically unlinked prevalence of subtype C infections can exceed 30% of the adult population (*7*). Therefore, this exploration of the implications of HIV variation for vaccine strain selection focuses on subtype C; we believe, however, that our reasoning and findings can be extrapolated to other intra- and intersubtype scenarios.

Although there is hope that a single vaccine strain may elicit a sufficiently cross-reactive response to confer a benefit, there is great interest in attempting to optimize vaccines through considerations of diversity. There are currently two general approaches to selecting vaccine strains that attempt to contend with the high levels of HIV sequence variation (Table 1). The first is based on using isolates of a particular subtype, sometimes selected from a geographic region where the vaccine is intended for use. Examples of this approach that are under way include the development of several A- and C-subtype vaccines, as well as CRF01 vaccine reagents (*5*, *6*). This kind of approach can be integrated with biological considerations, such as coreceptor usage, neutralization susceptibility, neutralization potency of the serum from the individual from whom the isolate was obtained, or the preferential use of isolates from recent seroconvertors (*8*). The second approach, rather than using actual viruses from within the population, is to construct either a consensus sequence or an ancestral sequence reconstructed on the basis of an evolutionary model. Such sequences have the advantage of being central and most similar to currently circulating strains of interest and may have enhanced potential for eliciting cross-reactive responses. They may also have economic and political advantages that merit consideration. Economic, because it is not feasible to duplicate vaccine design efforts using country-specific strains for every nation and region that needs a vaccine, and this is a way to limit the number of constructs that must be produced and tested in a way that is logical and scientifically defensible. Political, because such artificial sequences are not associated with any specific country of origin, so nations hosting vaccine trials would not need to contend with the natural concerns that arise when asked to host a vaccine trial using HIV-1 antigens with distant geographic origins. Some subtype C sequences of current interest are indicated in Fig. 1, a phylogenetic tree that includes geographic information to provide a basis for considering candidate vaccine strains. (Phylogenetic terms and concepts used in this article are defined in the legend to Fig. 1.) These two basic approaches, in different ways, directly confront the

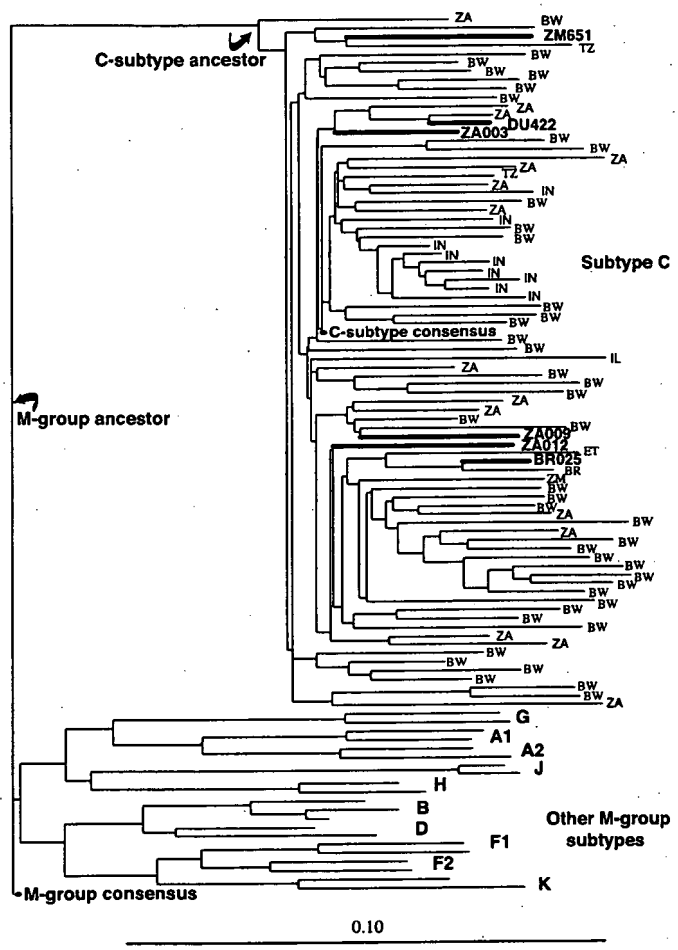

Fig. 1. Maximum likelihood phylogenetic tree showing the genetic distances and relationships of potential vaccine strains to subtype C *gag* sequences, and to representatives from other subtypes. The external nodes, the branch tips on the right of the tree, each represent an actual sequence. The interior nodes, or branch points, are ancestral to the "clade" of sequences that branch to their right. This tree uses the M-group consensus as the outgroup, a sequence brought into the analysis to help determine the ancestral states and root of the "ingroup," in this case the HIV-1 M group. Constructing the tree with the M-group consensus sequence as an outgroup forces the ancestral node of the M group to be central to all of the subtypes; using a more conventional strategy of selecting another primate lentiviral sequence for an outgroup can lead to statistically unsupported and unrealistic locations for ancestral nodes (*1*). The horizontal branch lengths in the trees represent evolutionary distances and indicate how many nucleotide substitutions have occurred; these are estimated from the evolutionary model. The two-letter code for the country of origin of subtype C sequences is indicated: India, IN; South Africa, ZA; Botswana, BW; Tanzania, TZ; Israel, IL; Ethiopia, ET; Zambia, ZM; and Brazil, BR. The locations of the C-subtype and M-group consensus and ancestor are indicated. Potential C-subtype vaccine strains are marked by a bold branch and their isolate name; these include Brazilian BR025, one of the first available subtype C isolates; ZM651 (AF286244), a Zambian strain; IN101 (AB023804), an Indian sequence discussed in the text, and sequences derived from recent South African isolates, including Du422 (AY043175), and three available for reagent development through the UNAIDS network, derived from viable isolates with full-length sequenced clones, ZA009 (AY118166), ZA003 (AY118165), and ZA012 (AF286227) (*24*, *53*). [Note: There are two South African samples designated ZA009, one from a B clade infection (AF095828), and one recently obtained UNAIDS Network C clade isolate; in this paper ZA009 refers to the C clade isolate.] The scale bar indicates the genetic distance along the branches. *gag* and *env* maximum likelihood trees with all sequences labeled are also available (*32*).

Table 1. A summary of classes of potential vaccine strains for use in subtype C epidemic regions. Differences show typical values of the percentage of amino acid changes observed when comparing the potential vaccine strain sequences to the sets of available C clade protein sequences. The lower bound represents conserved proteins, the upper bound variable proteins.

Vaccine source	Differences (%)	Advantages and characteristics
Isolates, subtype B	10–30	Furthest along in vaccine testing Based on an actual virus, and strains can be selected on the basis of advantageous biological characteristics
Isolates, subtype C	5–15	The closest natural form to C-subtype circulating strains Like subtype B isolates, based on actual viruses, and thus can be selected on the basis of biological features
Consensus, subtype C	3–8	Central to C-subtype circulating strains Each amino acid is most commonly found at that position
Ancestral, subtype C	3–8	Representative of the C subtype Maximum likelihood model of ancestor sequence
M-group consensus	5–15	Representative of the HIV-1 epidemic Most likely to cross-react with all clades Consensus of the subtype consensus sequences

problem of viral diversity; we will explore their advantages and disadvantages in the sections that follow.

Other concepts being explored to address diversity issues can ultimately be considered in the framework of the two basic approaches to strain selection described above. For example, multivalent cocktails of proteins that include a spectrum of regional variants are being evaluated. Most of these strategies assume that the immune responses elicited by any one circulating strain will be of sufficient cross-reactivity to protect against other strains from the same subtype. Given that intra-subtype diversity in variable proteins can reach 20%, even this assumption may be too optimistic. Consensus sequences and strains from viable isolates could be combined in a polyvalent approach. A second strategy is to design modified envelopes to enhance exposure of epitopes known to be capable of inducing broadly neutralizing antibodies (*9–11*). There are a limited number of monoclonal antibodies that have broad, cross-clade neutralization capabilities (*11–13*), and these antibodies can act synergistically (*13–15*). Vaccines specifically designed to target these conserved epitopes, if successful, may ultimately be optimized by fine-tuning as subtype-specific vaccines, as appropriate; although there is little evidence that subtypes correspond to neutralization phenotypes, in some cases particular clades can be refractive or less susceptible to particular antibodies. For example, the three broadly neutralizing monoclonal antibodies 2F5, 2G12, and IgG1b12, raised against B clade strains, were not individually able to neutralize a C clade primary isolate, although they could in combination (*14*). Even if subtypes are not relevant to a particular neutralizing epitope, lineage-specific variation within the relevant antigenic domain may still be worth considering. A third strategy for contending with diversity is to use polyvalent peptides spanning a region like the V3 loop that induces strain-specific neutralizing antibodies (*16*), to attempt to elicit a set of responses that together confer cross-reactive protection (*17, 18*).

Isolate-Based Vaccines

Geographic considerations. For historical reasons, AIDS vaccines reagents were first developed from subtype B viruses, the dominant subtype in the United States and Europe. It has been proposed that such strains be included in vaccine trials conducted in populations infected with subtypes other than B. Cytotoxic T lymphocyte (CTL) studies provide evidence for cross-subtype T cell responses, and B viruses have been studied for a longer time, so researchers can more rapidly move forward in safety and immunogenicity (phase I) studies (*5, 6*). However, T cell immune responses in general are more intense and have greater breadth within a subtype (*19–23*). Thus, although there is potential for cross-reactivity and even for synergistic interactions between antibodies (*15*), it is more than likely that both the breadth and intensity of polyclonal T and B cell immune responses to cross-clade immunogens will be suboptimal and that important epitopes will be missed. Subtype-specific, single-strain, or combination vaccines have been strongly advocated in recent years (*24–26*), and approximately 10 subtype C vaccines are poised to enter phase I trials in India, South Africa, and China (*27*).

There has been some discussion of choosing a regional strain for a vaccine, for example, an Indian strain to be used in India, and a South African strain in South Africa, and so on (*8*). There is little support for this in terms of sequence analysis. Subtype C sequences from Botswana and South Africa intermingle (*28*), and there is no obvious choice of a single sequence most representative of the diversity in these regional samples (*28*); however, selecting a sequence with a short branch length relative to the common ancestor (*29*) in C clade might be advantageous, as it would tend to be most similar to the majority of contemporary sequences represented in the tree (*30*). Conversely, it would be sensible to avoid selecting outliers. Indian sequences tend to form a distinct subclade (*31*) within the C clade, indicating that most sampled Indian viruses are descended from a single founder strain. A small number of sequences from African nations are associated with sequences from India, however, and the sampling is extremely limited relative to the scale of the epidemic in both regions. Thus, there may be continuing movement of the virus between Africa and India. The Indian clade sequences tend to have short branch lengths relative to the root of the C clade. As a consequence, a strain like IN101 (accession number AB023804) from India is closer to most African subtype C strains than African strains are to each other (*32*), an interesting quirk that emphasizes that it does not necessarily confer an advantage to select a strain from the country where a vaccine trial will be held.

There are other subclades within the C subtype, besides the Indian subclade (*33*), that could be considered as vaccine candidates (Fig. 1), but such subclades tend to have much shorter defining branch lengths than subtypes and, consequently, fewer distinguishing amino acids, so the benefit of considering them each separately for vaccines diminishes. On rare occasions, geographically localized epidemics have been identified soon after the introduction of a founder virus, and prevalent viruses were highly related, as in Thailand (*34*) and Kaliningrad (*35*). It may ultimately be advantageous to develop vectors and strategies for rapid-response vaccine programs in such circumstances, when a highly similar virus is spreading explosively through a vulnerable population, but first, a working vaccine concept must be in place.

Evolutionary evidence for subtype-specific antigenicity in the envelope protein. Clearly, subtype-specific vaccines would increase the overall sequence similarity of the vaccine antigen relative to circulating viruses (Fig. 2), but this is only part of the story for antibody binding, because protein folding and exposure of antigenic domains are of great importance. To explore the hypothesis that there may be subtype-specific patterns in the exposure of antigenic domains that are able to elicit antibody responses strong enough to drive escape

mutations, we compared estimates of codon-specific ratios of nonsynonymous to synonymous substitution rates (dN/dS) (*36*) in B clade and C clade envelope genes. (We selected B and C subtypes, as B clade vaccines are being considered for use in populations where the C subtype dominates.) High rates of diversifying selection were identified in different regions of the envelope protein (Env) in the two lineages, most strikingly in the Env V3 to C4 region. The V3 loop is less variable in the C subtype than in other subtypes (*37*), and as expected, the density of sites in the V3 loop with dN/dS > 1 was higher in the B clade than in the C clade. This pattern was reversed, however, in the region just proximal to the V3 loop, where multiple sites show an excess of nonsynonymous substitutions in the C clade but not in the B clade (Fig. 3). To explore the consistency of these patterns within the C clade, three subclades, or phylogenetically associated groups of sequences within the C subtype, were examined independently. Two sets had 21 subtype C sequences, and the third had 18 sequences. The results suggested substantial intraclade coherence in how selection acts on individual codons within the C subtype; the 12 strongly selected sites in the region downstream of the V3 loop (Fig. 3) had a dN/dS ratio > 1 in each of the three independent C-subtype data sets, and the tip of the V3 loop was relatively constrained with low dN/dS values. Given that immune escape is likely to be a driving force of positive selection, immune pressure may be focused on different regions of Env in the B subtype (the V3 loop) and C subtype (the COOH-terminal region beyond the V3 loop). If this interpretation of the observed differences in selection pressure in B and C subtypes is correct, there may be advantages in using a clade-appropriate vaccine strain, as the immune response to the vaccine and the circulating virus would share antigenic domains.

Artificial Sequences for Minimizing Diversity

An effective way to minimize the degree of sequence dissimilarity between a vaccine strain and contemporary circulating viruses is to create artificial sequences that are "central" to these viruses. The simplest way to design such a sequence is to use a consensus sequence based on the most common amino acid in each position in an alignment (*33*, *38*). Alternatively, a model of the most recent common ancestral sequence of an appropriate lineage can be reconstructed from a phylogenetic tree, for example, by means of maximum likelihood. The most likely sequence at any interior node in a tree can be derived from the sequences used to construct the tree, the evolutionary model used (how often one base is mutated to another, and the relative mutation rate at each site), and the branching pattern of the tree. Figure 1 illustrates where the C consensus and ancestral branch points are located in the tree. Both of these sequences are more "central," i.e., they are closer to modern C-subtype sequences than modern sequences are to each other. As artificial sequences, their construction depends on the sequences included in the analysis and so will change as the database expands.

Envelope proteins are the most difficult HIV proteins to construct artificially, as both ancestral and consensus sequences contain hypervariable domains with multiple insertions and deletions (indels). Alignments are subjective in such regions, and indels do not evolve according to the base substitution models currently assumed in deriving a maximum likelihood tree. For constructing our consensus and ancestor sequences (*3*), hypervariable regions are aligned by anchoring on glycosylation sites, and only minimal common elements spanning the region are retained. As both consensus and ancestral sequences are derived and not actual sequences, expression, antigenicity, and biological activity require careful characterization before use in a vaccine (*39*).

Although artificial sequences may not have a proper protein conformation, and this may be critical for antibody responses, it is less important for designing T cell epitopes

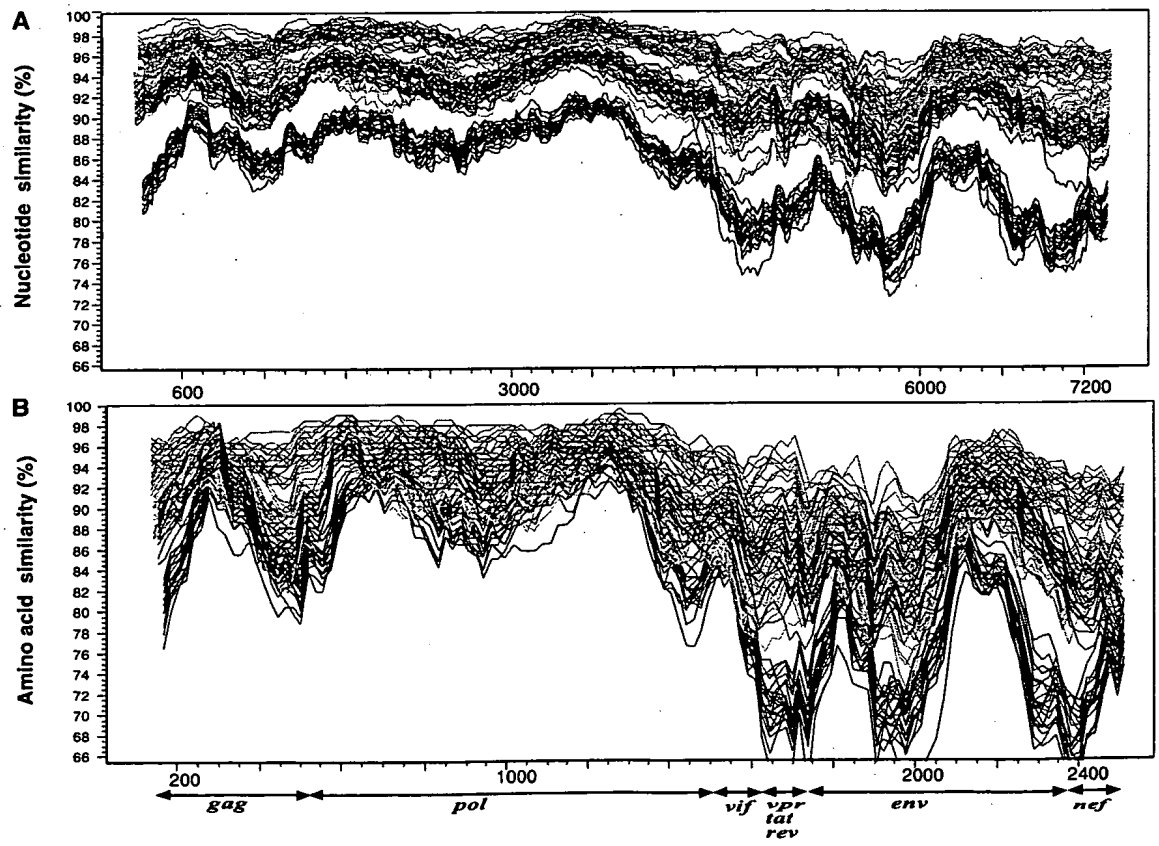

Fig. 2. Scanning the HIV-1 genome and proteins to illustrate similarities between potential vaccine candidates and sequences from isolates. (**A**) and (**B**) compare 23 full-length subtype C sequences from South Africa, Botswana, and India with potential vaccine sequences. Green lines represent the comparison with the subtype C consensus sequence. The purple and blue lines show the comparison of the sequences of vaccine candidates BR025 and ZA003 (described in Fig. 1), respectively. The red lines show an interclade comparison of subtype C sequences with the B clade sequence JRCSF. (A) shows a nucleotide similarity plot, and (B) shows the corresponding amino acid similarity plot.

or peptide reagents for testing T cell responses. Consensus sequences may be ideal for peptides used to explore the T cell immune response, as it would probably improve recognition compared with any single reference strain, and using sets of autologous strain peptides can be prohibitively expensive. A consensus may even be preferable to autologous peptides, as CTL escape mutations can rapidly predominate in the viral quasispecies, and important early responses (*40*) may go undetected through the use of peptides based on isolates from later time points that have escaped the early responses. Consensus peptides for several subtypes are available (*41*).

A similarity plot maps the percent similarity of a query sequence relative to a test set in a window spanning a region of a specified size that is moved progressively along an alignment. In Fig. 2, prototype vaccine reagents (a C consensus, two subtype C vaccine strains, and a subtype B isolate) are used as query sequences and compared with 23 subtype C sequences from South Africa, India, and Botswana. In every gene region, the same relative pattern holds. The spectrum of similarity scores for the C consensus sequence compared with the set of 23 C sequences is 5 to 15% greater than when any one C isolate is compared with others in the set (*28*). In turn, subtype C proteins are 5 to 15% more similar to the subtype C sequences than are subtype B sequences. This implies that using a B clade virus as the basis of a vaccine in a C clade–dominated epidemic may be less effective than using a C clade virus, and a C clade virus may not be as effective as a C consensus. Conserved proteins from different subtypes can be more closely related than variable proteins from the same subtype, and this fact might be exploited by using a single vaccine strain for conserved proteins and multiple clade-specific strains for variable proteins.

We have been discussing pooling sequences within a subtype to generate artificial central sequences, but it is also possible to pool the subtypes themselves. To maximize potential cross-reactivity, we have created sequences central to the M group, the diverse viruses that have contributed most to the global epidemic. The set of subtype consensus sequences was used to build an M-group consensus, thus weighting the subtypes equally. The M-group consensus and the most recent common ancestor can be very nearly identical (*1*, *28*). Because of the nature of the HIV-1 M-group phylogeny, the average distance from HIV-1 sequences to the M-group consensus is similar to intrasubtype sequence distances between contemporary isolates (*32*), roughly half that of intersubtype distances (Table 2). In the Democratic Republic of the Congo (DRC) (*42*, *43*), so many subtypes and recombinants circulate together that the extent of the regional diversity resembles the global diversity. In this setting, an M-group consensus may be helpful, or a polyvalent approach including representative strains from common subtypes along with the M-group consensus. Even in a design focusing on epitopes that are conserved across clades, an M-group consensus might be the optimal baseline sequence. Consensus and ancestral sequences for the major HIV-1 subtypes, CRFs, and the M group are available (*3*) and will be updated as sequences accrue. Intersubtype similarity comparisons with an M-group consensus are included in the supplementary material (*31*).

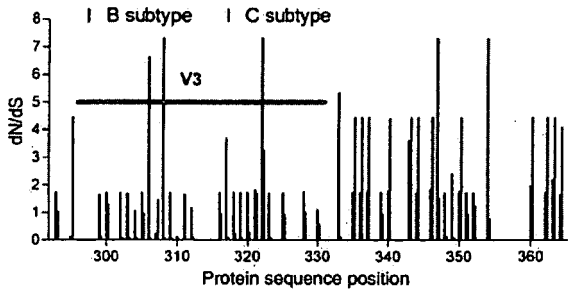

Fig. 3. The dN/dS ratio at each position in the V3 region, comparing a B-subtype and a C-subtype alignment. The dN/dS ratio was determined for each codon in an *env* alignment of C- and B-subtype sequences. The V3 region gave a particularly striking distinction between the two subtypes, illustrated here. The blue lines indicate the dN/dS ratio in the V3 loop of the B subtype, a region known to be a target of type-specific neutralizing antibodies. Four codons on either side of the tip of the V3 loop have dN/dS ratios over 5, indicative of very strong positive selection. The red lines indicate the dN/dS ratio of the C subtype, and there is no strong pressure for change near the tip of the V3 loop. In contrast, downstream of the V3 loop there are 12 codons that exhibit high dN/dS ratios (>4) in the C subtype and only three in the B subtype. This suggests different regional evolutionary pressures in the two subtypes, and possibly distinct regions of antigenic exposure in these regions in the B and C lineages.

Consensus and ancestral sequences conserve CTL epitopes. Experimentally defined CTL epitopes in the HIV Immunology Database (*3*) cluster more densely in conserved regions of HIV proteins (*44*). The peptides spanning variable regions used to detect CTL responses can be quite different from the infecting strain that elicited the response, no doubt contributing to the paucity of defined epitopes in variable domains, but, in addition, an enrichment of features that could contribute to CTL escape can be discerned in variable domains (*45*). Either way, regions where defined epitopes are concentrated are likely to be key for cross-reactive CTL responses (*44*). The epitopes in the database have primarily been defined for B clade responses; however, the C clade peptides that trigger immunodominant responses tend to be localized in these same regions (*44*, *46*). Thus, in contrast to Fig. 2 and Table 2, where whole proteins were analyzed, we focused on protein regions where CTL epitopes have been found in order to create Fig. 4, which shows the average sequence distances from potential vaccine strains to immunogenic regions in subtype C proteins, by country of origin. Three proteins were selected, representing the spectrum of variability: highly conserved p24, variable p17, and highly variable envelope (subunit gp160). In the immunogenic regions analyzed, C-subtype consensus and ancestral sequences had the fewest amino acid changes relative to contemporary C-subtype protein sequences. Within-subtype comparisons of single C-subtype viral strains and the M-group consensus sequences gave comparable numbers of amino acids changes, roughly half the number of changes relative to B subtype interclade comparisons.

Consensus and ancestral sequences conserve predicted immunoproteasome cleavage

Table 2. Median and range of percent similarity scores between potential vaccine candidates and an alignment of C clade sequences. The similarities are shown for the p24 and gp160 proteins, representative of highly conserved and highly variable proteins. The C-subtype ancestral and consensus (con) sequences, when compared with the set of protein sequences from contemporary subtype C isolates, have comparable distributions of similarity scores. The M-group consensus is comparable to C clade isolate sequences. Two representative C clade isolates are shown, DU422 from South Africa and BR025 from Brazil. The last column shows the results of a subtype B strain compared with C sequences (intersubtype) for contrast. Regions with gaps were included, and either a relative insertion or deletion or an amino acid change at any position is considered one difference.

Protein	Consensus sequences			Isolate sequences		
	C-subtype con	C ancestral	M-group con	C_ZA.DU422	C_BR.92BR025	B_FR.HXB2R
p24	95.4(97.4–92.8)	94.8(97.7–92.8)	93.1(95.1–91.1)	94.8(97.4–92.1)	92.8(96.4–88.9)	88.9(91.8–86.6)
gp160	87.5(90.6–83.5)	86.1(88.3–81.7)	81.4(84.0–77.1)	81.3(89.3–77.4)	83.1(85.9–79.6)	72.1(73.8–68.0)

sites. For viral proteins to be recognized by CTL they must be processed, and each step of epitope processing has potential constraints imposed by sequence specificity. Immune escape due to mutations in epitope flanking regions demonstrates escape from immune suppression through cleavage abrogation (*47*) and shows that epitope processing is sensitive to the surrounding sequence, although a simple cleavage signal is not readily discernable. If the tendency to be cleaved at a relevant site is markedly different in a vaccine strain and a challenge strain, the immunological priming induced by the vaccine will be ineffective. This problem is difficult to resolve experimentally, so we addressed it computationally by means of NetChop (*48*), a neural net prediction program for immunoproteasome cleavage (*32*, *45*).

The median cleavage prediction scores for subtypes B and C were correlated, but although many sites preserved their relative tendency to be cleaved, there were many exceptions, positions with high cleavage prediction scores in subtype B but not in C, or vice versa (*32*). This suggests that the predilection for cleavage of many sites would be altered in the two subtypes, which could result in diminished breadth of cross-reactive responses. C clade sequences and the M-group consensus gave cleavage prediction patterns that were similar when compared with the median scores for the C clade alignment, and they performed better than sequences from the B clade (*32*). Scores predicted for the C-subtype consensus cleavage correlated most strongly with the median scores for the subtype C population (*32*), suggesting that it would be processed at any given position similarly to most of the subtype C strains, and so it may have the greatest potential for eliciting cross-reactive immune responses at the population level. The complete analysis is provided in the supplemental information (*32*), but in summary, the linear correlation coefficients (r^2 values) for comparisons of the median C clade cleavage scores to vaccine candidate strains are as follows for positions in the Envelope protein: B clade, 0.65; the M-group consensus, 0.79; the M-group ancestor, 0.80; specific sequences from subtype C isolates, 0.79 to 0.81; the subtype C ancestor, 0.88; and the subtype C consensus, 0.92.

Consensus or reconstructed ancestor? One might assume that an ancestral sequence would resemble more closely a real viral protein than a consensus. It is statistically extremely unlikely, however, that an ancestor corresponds to an ancestral sequence of a clade as complex and diverse as an HIV-1 subtype. Furthermore, reconstruction greatly depends on assumptions inherent in building maximum likelihood trees; for example, if positions are not evolving independently or there are undetected recombination events, the ancestral reconstruction would be influenced and incorrect. Thus, it is highly improbable that an ancestor of a subtype ever existed precisely as reconstructed. Ancestor and consensus sequences are subject to different sampling biases and will change from year to year as sequences accrue. An ancestor is influenced by sequences external to the subtype of interest and will tend to be slightly more distant from available sequences within a subtype than a consensus sequence (Table 2), as well as slightly closer to sequences of other subtypes. The inclusion of a new outlier that branches near the basal node of a subtype could have a strong influence on the ancestral node (see, for example, Fig. 1, where the ancestral and consensus sequences are separated mainly because of one single outlier virus from South Africa), but as a single sequence it would have little bearing on the consensus. In contrast, a consensus will be influenced by the sampling of sequences from within subclades. For example, if many sequences were obtained from the Indian subclade during the next year, the next C consensus would be more like the Indian subset, but the shift in sampling would have less impact on the subtype C ancestor, unless the new sequences substantially altered the evolutionary model.

It is possible that a consensus sequence based on contemporary isolates may be more likely to reflect escape variants relevant to the host population than an ancestral sequence. For example, a CTL escape mutant in an epitope presented by a human leukocyte antigen molecule common in a certain population may be selected for and may be more likely to be represented in the consensus sequence than a reconstructed ancestor sequence. If most viruses in the circulating population had already lost the original epitope because of immune escape, and if the epitope elicited a dominant response upon vaccination with a strain that carried it, then the consensus sequence would have an advantage. On the other hand, if the wild-type form of the epitope was still circulating, even infrequently, and if the epitope was particularly potent, there might be an advantage in using the ancestor. In the end, both concepts need to be tested experimentally, both in terms of B cell and T cell responses.

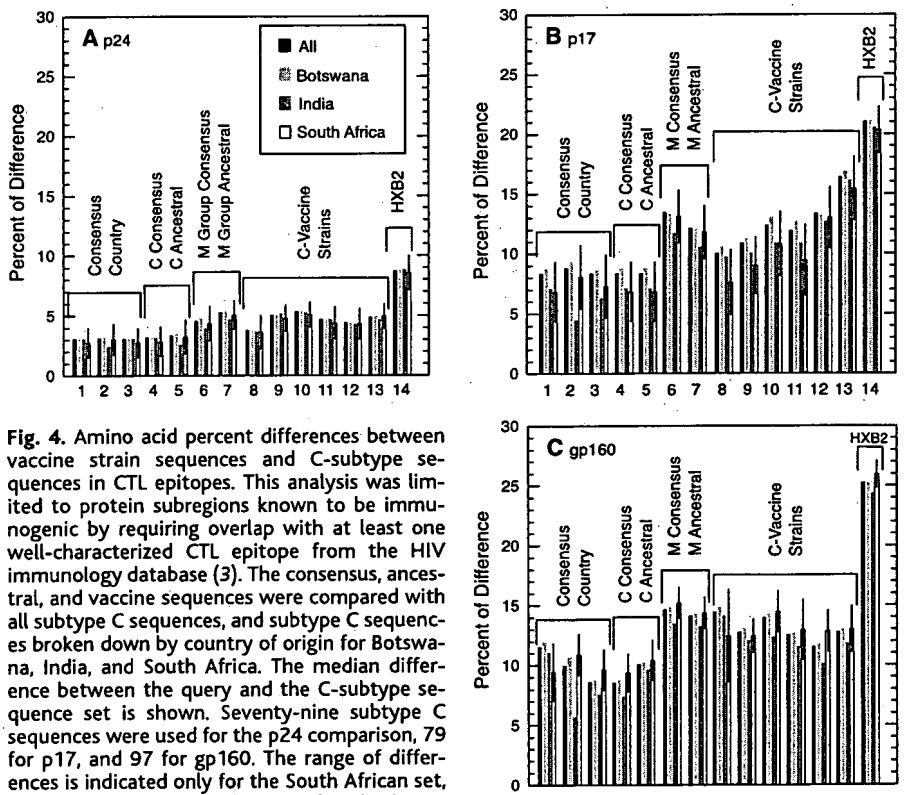

Fig. 4. Amino acid percent differences between vaccine strain sequences and C-subtype sequences in CTL epitopes. This analysis was limited to protein subregions known to be immunogenic by requiring overlap with at least one well-characterized CTL epitope from the HIV immunology database (*3*). The consensus, ancestral, and vaccine sequences were compared with all subtype C sequences, and subtype C sequences broken down by country of origin for Botswana, India, and South Africa. The median difference between the query and the C-subtype sequence set is shown. Seventy-nine subtype C sequences were used for the p24 comparison, 79 for p17, and 97 for gp160. The range of differences is indicated only for the South African set, which tends to be typical, to simplify the figure. The comparisons are numbered along the *x* axis: 1, Botswanan C consensus; 2, Indian C consensus; 3, South African C consensus; 4, C clade consensus; 5, C ancestral sequence; 6, M-group consensus; 7, M-group ancestral sequence; 8, C.ZA.DU422; 9, C.ZA.ZA003; 10, C.ZA.ZA009; 11, C.ZA.ZA012; 12, C.ZM.ZM651; 13, C.BR.BR025; and 14, B.FR.HXB2R.

Applying Evolutionary Principles to Vaccine Strain Selection

How can HIV's evolutionary trajectory be incorporated into a sensible vaccine approach? Although subtypes of HIV-1 are

phylogenetically defined on the basis of genetic and evolutionary distances, the practical consequence of phylogenetic clustering of viruses is patterns of shared amino acids that can influence the immunological cross-reactivity of vaccine-stimulated immune responses. Env proteins from different clades can differ in more than 30% of their amino acids, and HIV-1 continues to diversify. Neutralizing antibody as well as CTL escape occurs in vivo (*49*, *50*), escape mutations can be transmitted and stable (*49*), and there are protein regions under clear positive selection pressure (*36*) (Fig. 3). These observations indicate that HIV-1 amino acid variation is immunologically relevant. The impact of that variation on vaccine-conferred immune protection will ultimately have to be assessed through vaccine trials, but the differences between potential vaccines and circulating strains can be minimized when designing trial reagents to attempt to enhance cross-reactive responses.

Most vaccines are intended to elicit polyclonal responses to multiple epitopes, so even if they differ in some antigenic domains from a given virus, in others they may be cross-reactive. Selecting a clade-appropriate vaccine for a regional trial would tend to increase the number of potentially cross-reactive epitopes by increasing the level of similarity between the vaccine and the population, and the use of consensus and ancestors would enhance the cross-reactive potential. The difference in selection pressure on B and C clade envelopes is indicative of lineage-specific antigenicity, further supporting the use of subtype-appropriate vaccines to maximize the probability that the vaccine elicits immune responses to domains that are antigenic in the circulating viruses.

We could see no compelling advantage in further subdividing the C clade by country of origin, although this is often a consideration for vaccine design (*8*). Our analysis supports the recommendations of the international meeting on candidate vaccines for the developing world sponsored by the Vaccine Research Center of the United States, National Institute of Allergy and Infectious Diseases (*51*), indicating that, although there may be advantages to a subtype-specific vaccine, a promising subtype-specific vaccine candidate could be used in many different geographic locations without compromising the potential for success. This does not mean that there would never be an advantage in tailoring a vaccine further by selecting a sequence from an interior subclade within a subtype. For example, there might be an advantage in using an Asian, not African, CRF01 in Thailand, or an Indian C clade sequence in India. But within-clade differences tend to be subtle and represent far fewer amino acid changes than between-subtype differences.

In regions where an epidemic is dominated either by a particular subtype or CRF, it makes sense to use that dominant lineage for a vaccine and to consider the use of a consensus or ancestor. Although we cannot know if even the use of central sequences will be enough to contend with HIV diversity, this kind of strategy can potentially enhance the cross-reactivity and breadth of a vaccine response relative to any single strain. In regions where two or three subtypes and multiple recombinants are cocirculating, to include each of the prevalent subtypes could improve the potential coverage not only of those subtypes, but of the variety of recombinant forms that stem from them (*52*). Finally, nations with very diverse viral populations, like the DRC, might be best served by developing polyvalent vaccines including a spectrum of natural forms combined with an M-group consensus. An M-group consensus or ancestor is central not only to the major subtypes, but to recombinant forms involving the subtypes. Even if a single subtype predominates in a country, combining an M-group consensus with a regionally dominant subtype might be advantageous in an urban context where people of many nationalities mingle.

References and Notes

1. B. Korber *et al.*, *Science* **288**, 1789 (2000).
2. P. M. Sharp, E. Bailes, D. L. Robertson, F. Gao, B. H. Hahn, *Biol. Bull.* **196**, 338 (1999).
3. HIV Immunology and Sequence Databases, B. Korber *et al.* Eds. (Los Alamos National Laboratory, Los Alamos, NM, 2000); available at www.hiv.lanl.gov.
4. B. T. Korber, B. Foley, B. Gaschen, C. Kuiken, in *Retroviral Immune Response and Restoration*, G. Pantaleo and B. D. Walker, Eds. (Humana Press, Totowa, NJ, 2001), pp. 1–32.
5. A. M. Schultz, J. A. Bradac, *AIDS* (*London*) **15** (suppl. 5), S147 (2001).
6. B. Graham, in *HIV Molecular Immunology*, B. T. Korber *et al.*, Eds. (Los Alamos National Laboratory, Theoretical Biology, Los Alamos, NM, 2000), Part I, pp. 20–38.
7. UNAIDS, www.unaids.org/
8. J. Goudsmit, in *IAVI Rep. Dec 2000/Jan 2001* (International AIDS Vaccine Initiative, New York, 2001).
9. S. W. Barnett *et al.*, *J. Virol.* **75**, 5526 (2001).
10. J. M. Binley *et al.*, *J. Virol.* **74**, 627 (2000).
11. E. O. Saphire *et al.*, *Science* **293**, 1155 (2001).
12. W. Xu *et al.*, *J. Hum. Virol.* **4**, 55 (2001).
13. M. B. Zwick *et al.*, *J. Virol.* **75**, 12198 (2001).
14. J. R. Mascola *et al.*, *J. Virol.* **73**, 4009 (1999).
15. A. Li *et al.*, *J. Virol.* **72**, 3235 (1998).
16. H.-X. Liao *et al.*, *J. Virol.* **74**, 254 (2002).
17. B. F. Haynes *et al.*, *AIDS Res. Hum. Retrovir.* **11**, 211 (1995).
18. For example, there were 34 unique forms of the immunogenic tip of the V3 loop (corresponding to the IHIGPGRA of MN) among C clade sequences from 436 infected Africans in the HIV Sequence Database 2001, and these may fall into a smaller number of workable serotypes [S. Zolla-Pazner, M. K. Gorny, P. N. Nyambi, T. C. VanCott, A. Nadas, *J. Virol.* **73**, 4042 (1999)] that could serve as a basis for a polyvalent peptide, but the complexity of this problem rapidly increases when multiple subtypes are considered.
19. H. Cao *et al.*, *J. Infect. Dis.* **182**, 1350 (2000).
20. L. Dorrell *et al.*, *Eur. J. Immunol.* **31**, 1747 (2001).
21. G. Ferrari *et al.*, *Immunol. Lett.* **79**, 37 (2001).
22. V. Novitsky *et al.*, *J. Virol.* **75**, 9210 (2001).
23. S. L. Rowland-Jones *et al.*, *J. Clin. Invest.* **102**, 1758 (1998).
24. J. van Harmelen *et al.*, *AIDS Res. Hum. Retrovir.* **17**, 1527 (2001).
25. S. A. Lee *et al.*, *Vaccine* **20**, 563 (2001).
26. D. P. Francis *et al.*, *AIDS Res. Hum. Retrovir.* **14** (suppl. 3), S325 (1998).
27. K. Gupta, personal communication, International AIDS Vaccine Initiative (IAVI).
28. M. Groenink *et al.*, *Science* **260**, 1513 (1993).
29. E. B. Stephens *et al.*, *J. Med. Primatol.* **25**, 175 (1996).
30. B. Foley, H. Pan, S. Buchbinder, E. L. Delwart, *AIDS Res. Hum. Retrovir.* **16**, 1463 (2000).
31. R. Shankarappa *et al.*, *J. Virol.* **75**, 10479 (2001).
32. Supplemental materials available on *Science* Online concerning methods, detailed figures, and additional discussion include *gag* and *env* trees used in this paper; interclade similarity plots; immunoproteasome cleavage prediction comparisons; and consensus and ancestral sequences for major subtypes, CRFs, and the M group.
33. V. Novitsky *et al.*, *J. Virol.* **76**, 5435 (2002).
34. F. E. McCutchan *et al.*, *J. Virol.* **70**, 3331 (1996).
35. K. Liitsola *et al.*, *AIDS* (*London*) **12**, 1907 (1998).
36. Z. Yang, R. Nielsen, N. Goldman, A. M. Pedersen, *Genetics* **155**, 431 (2000).
37. C. L. Kuiken, B. Foley, E. Guzman, B. T. Korber, in *Molecular Evolution of HIV*, K. Crandall, Ed. (Johns Hopkins Univ. Press, Baltimore, MD, 1999).
38. B. Korber *et al.*, *Br. Med. Bull.* **58**, 19 (2001).
39. An M-group consensus envelope protein reacted equally well with sera from B and subtype C infections in Western blots, and BiaCore assays revealed that it bound to CD4 and numerous monoclonal antibodies, equivalently to a normal Env; further studies are under way. F. Gao and B. Hahn, unpublished data (2002).
40. T. M. Allen *et al.*, *Nature* **407**, 386 (2000).
41. The NIH AIDS Reagent Program www.aidsreagent.org/.
42. J. L. Mokili *et al.*, *AIDS Res. Hum. Retrovir.* **15**, 655 (1999).
43. N. Vidal *et al.*, *J. Virol.* **74**, 10498 (2000).
44. B. D. Walker, B. T. Korber, *Nature Immunol.* **2**, 473 (2001).
45. K. Yusim *et al.*, *J. Virol.*, in press.
46. P. J. Goulder *et al.*, *J. Virol.* **74**, 5679 (2000).
47. N. J. Beekman *et al.*, *J. Immunol.* **164**, 1898 (2000).
48. C. Kesimir, A. K. Nussbaum, H. Schild, V. Detours, S. Brunak, *Protein Eng.* **15**, 287 (2002).
49. P. J. Goulder *et al.*, *Nature* **412**, 334 (2001).
50. J. P. Langedijk, G. Zwart, J. Goudsmit, R. H. Meloen, *AIDS Res. Hum. Retrovir.* **11**, 1153 (1995).
51. NIAID/NIH report: Development of URC HIV Candidate Vaccines for the Developing World, www.vrc.nih.gov.
52. S. M. Agwale *et al.*, *Vaccine* **20**, 2131 (2002).
53. C. M. Rodenburg *et al.*, *AIDS Res. Hum. Retrovir.* **17**, 161 (2001).

54. We thank Y. Li, Y. Chen, and C. M. Rodenburg for excellent technical assistance and C. Brander, J. Bradac, C. Kuiken, U. Smith, and J. Mullins for ideas and suggestions. We would also like to thank our reviewers and editor, B. Jasny, for their exceptionally detailed and thoughtful comments. This work was supported by grants from the National Institutes of Health (N01 AI 85338, P20 AI 27767, R01 AI 40951, R01 AI 35351, R01 AI 05397) and NIH–Department of Energy interagency agreement YI AI 1500-01 and internal Laboratory Directed Research and Development funding at Los Alamos National Laboratory.

Supporting Online Material

www.sciencemag.org/cgi/content/full/296/5577/2354/DC1

Materials and Methods

Figs. S1 to S3